
**دانشگاه تهران**

**پردیس دانشکده های فنی**

**دانشکده مهندسی برق و کامپیوتر**


# گزارش کارآموزی

**عنوان کارآموزی: هوش مصنوعی و تحلیل داده**

**نام محل کارآموزی: شرکت مهندسی صنایع یاس ارغوانی**


**نام و نام خانوادگی دانشجو: محمد امانلو**

**شماره دانشجویی: ۸۱۰۱۰۰۰۸۴**


**نام استاد کارآموزی: دکتر رشاد حسینی**

**تاریخ انجام کارآموزی: ۱۴۰۳/۴/۱۶**

## فهرست مطالب

## چکیده

این گزارش شامل شرح کاملی از فعالیت‌ها و تجربیات کسب شده در دوره کارآموزی در شرکت "مهندسی صنایع یاس ارغوانی" است که در حوزه علم داده و مدیریت خدمات انجام شد. در این دوره، با استفاده از داده‌های بزرگ بانک ملت، پروژه‌های مختلفی در زمینه‌های ارتقا مدل امتیازدهی پذیرندگان، پیش‌بینی رفتار مشتریان به منظور ارائه یک سیستم هشدار سریع، تحلیل حواله‌های کارت به کارت، شناسایی رفتارهای نامتعارف تراکنش‌ها، کاندید نمودن پایانه‌های خودپرداز مستعد با استفاده از بهینه‌سازی توزیع و استقرار هوشمند شعب و دستگاه‌های خودپرداز و پیش‌بینی منابع و نقدینگی مشتریان بانک ملت انجام شد. همچنین، طراحی و پیاده‌سازی مدیریت خدمات بر اساس ITIL 4 با استفاده از BPMN و سیستم‌های Wendia/Jira نیز از دیگر وظایف مهم در این دوره بود. در این گزارش دو پروژه اصلی که در این تابستان به پایان رسید یعنی پروژه پیش‌بینی رفتار مشتریان به منظور ارائه یک سیستم هشدار سریع و پروژه کاندید نمودن پایانه‌های خودپرداز مستعد با استفاده از بهینه‌سازی توزیع و استقرار هوشمند شعب و دستگاه‌های خودپرداز به طور کامل شرح داده شده است.


# فصل اول

# معرفی محل کارآموزی

### ۱-۱- مقدمه

کارآموزی خود را در شرکت "مهندسی صنایع یاس ارغوانی" انجام دادم. این شرکت به عنوان یکی از پیشروان در ارائه خدمات مهندسی و مشاوره در حوزه‌های مختلف صنعتی شناخته می‌شود و در زمینه‌های متنوعی مانند مدیریت خدمات، علم داده، و تحلیل داده‌های بزرگ فعالیت دارد. با توجه به تحولات سریع در دنیای فناوری و نیاز روزافزون صنایع به داده‌ها و تحلیل‌های دقیق، شرکت یاس ارغوانی توانسته است خود را به عنوان یک بازیگر کلیدی در این حوزه معرفی کند.

شرکت یاس ارغوانی به دلیل رویکرد نوآورانه و استفاده از تکنولوژی‌های پیشرفته، به عنوان یک مرجع معتبر در صنعت شناخته می‌شود. این شرکت با بهره‌گیری از تیمی متشکل از متخصصان و کارشناسان با تجربه، توانسته است پروژه‌های متعددی را در زمینه‌های مختلف به انجام برساند. این تیم شامل مهندسان نرم‌افزار، تحلیلگران داده، مشاوران مدیریت و متخصصان حوزه‌های مختلف صنعتی است که هر یک با تخصص‌های منحصر به فرد خود به بهبود فرآیندها و ارائه راه‌حل‌های کارآمد کمک می‌کنند.

یکی از مهم‌ترین پروژه‌های این شرکت، همکاری با بانک ملت برای بهینه‌سازی فرآیندهای بانکی و تحلیل داده‌های بزرگ این بانک بود. این پروژه به دلیل حجم بالای داده‌ها و پیچیدگی فرآیندهای بانکی، چالشی بزرگ برای تیم یاس ارغوانی به شمار می‌رفت. هدف اصلی این پروژه، شناسایی الگوهای رفتاری مشتریان و بهینه‌سازی

خدمات بانکی بر اساس این الگوها بود. این کار شامل تحلیل داده‌های تراکنش‌ها، بررسی رفتار مشتریان و شناسایی نقاط قوت و ضعف در فرآیندهای موجود بود.

در این پروژه، تیم یاس ارغوانی از ابزارهای پیشرفته علم داده و تحلیل داده‌های بزرگ استفاده کرد. با استفاده از تکنیک‌های یادگیری ماشین و تحلیل داده، توانستند به نتایج قابل توجهی دست یابند. این نتایج به بانک ملت کمک کرد تا فرآیندهای خود را بهبود بخشد و خدمات بهتری به مشتریان ارائه دهد. همچنین، این پروژه به بانک ملت این امکان را داد که به تحلیل‌های دقیق‌تری از داده‌های خود دست یابد و تصمیمات بهتری در زمینه استراتژی‌های تجاری اتخاذ کند.

شرکت یاس ارغوانی همچنین در زمینه‌های دیگری نظیر بهینه‌سازی زنجیره تأمین، مدیریت پروژه، و مشاوره در زمینه فناوری اطلاعات نیز فعالیت دارد. این شرکت با ارائه راه‌حل‌های تخصصی و مشاوره‌های کارآمد، به سازمان‌ها کمک می‌کند تا عملکرد بهتری داشته باشند و به اهداف خود دست یابند. به عنوان مثال، در پروژه‌های بهینه‌سازی زنجیره تأمین، تیم یاس ارغوانی به بررسی فرآیندهای موجود و شناسایی نقاط ضعف می‌پردازد و با ارائه راهکارهای بهینه، به افزایش کارایی و کاهش هزینه‌ها کمک می‌کند.

در نتیجه، کارآموزی من در شرکت "مهندسی صنایع یاس ارغوانی" تجربه‌ای بسیار ارزشمند بود. این تجربه به من این امکان را داد که با محیط حرفه‌ای کار آشنا شوم و از نزدیک با پروژه‌های بزرگ و چالش‌های موجود در صنعت آشنا شوم. همچنین، از طریق همکاری با تیم‌های مختلف، توانستم مهارت‌های فنی و مدیریتی خود را تقویت کنم و به درک عمیق‌تری از مفاهیم علم داده و تحلیل داده‌های بزرگ دست یابم. این تجربه نه تنها به من کمک کرد تا دانش خود را گسترش دهم، بلکه به من آموخت که چگونه می‌توان با کار گروهی و همکاری مؤثر به نتایج بهتری دست یافت.

## ۲-۱- معرفی کلی و اطلاعات عمومی شرکت.

| |
|---|
| نام شرکت/موسسه: شرکت مهندسی صنایع یاس ارغوانی |
| آدرس: آدرس: تهران، خیابان پاسداران، خیابان گیلان شرقی، نبش خیابان شهید عراقی، پلاک ۵۷، طبقه ۵<br>تلفن: تلفن: ۰۲۱-۲۷۳۱۸۰۰۱<br>وب سایت: https://www.yaasie.com/<br>فکس: ۰۲۱-۲۲۵۷۰۱۹۳۲<br>شرکت مهندسی صنایع یاس ارغوانی |
| نوع محل کارآموزی: شرکت خصولتی/ دانش بنیان/ شرکت مهندسی |

| |
|---|
| تعداد تقریبی کارمندان شرکت: ۷۲ نفر حاضر در محل و تعدادی خارج از محل |
| حوزه فعالیت: تحلیل داده، تولید نرم‌افزار، هوش کسب‌وکار، مدیریت خدمات و مشاوره مدیریت |
| عنوان کارآموزی دانشجو: تحلیل داده و هوش مصنوعی |

## فصل دوم

# فعالیت های انجام شده توسط دانشجو در دوره کارآموزی

### ۲-۱- مقدمه

در این فصل، به شرح دقیق فعالیت‌هایی که در دوره کارآموزی انجام داده‌ام، پرداخته می‌شود. این فعالیت‌ها شامل طراحی و پیاده‌سازی مدیریت خدمات بر اساس ITIL 4، مشارکت در پروژه‌های مرتبط با داده‌های بزرگ بانک ملت، و تکمیل یک دوره آموزشی در زمینه یادگیری ماشین بود. هدف از این فعالیت‌ها، کسب تجربه عملی و افزایش دانش و مهارت‌های فنی در حوزه علم داده و مدیریت خدمات بود.

### ۲-۲- پروژه‌های مرتبط با داده‌های بانک ملت:

### ۲-۲-۱ کاندید نمودن پایانه‌های خودپرداز مستعد با بهینه‌سازی توزیع و استقرار هوشمند شعب و دستگاه‌های خودپرداز

یکی از پروژه‌های مهم، بهینه‌سازی توزیع شعب و دستگاه‌های خودپرداز بانک ملت بود. در این پروژه، با تحلیل داده‌های تراکنش‌های بانکی، مدل‌هایی برای بهینه‌سازی مکان‌یابی شعب و دستگاه‌های خودپرداز توسعه داده شد. هدف اصلی این پروژه، کاهش هزینه‌های عملیاتی و افزایش رضایت مشتریان با دسترسی آسان‌تر به خدمات بانکی بود. سپس در این پروژه، پایانه‌های خودپرداز مستعد جهت جایگزینی با پایانه‌های بانکی غیر نقد شناسایی شدند. با تحلیل داده‌های عملکرد و استفاده از الگوریتم‌های بهینه‌سازی، پایانه‌هایی که نیاز به بهبود یا جایگزینی داشتند، تعیین شدند.

| **بیان مسأله پروژه** |
|---|
| در این پروژه سعی بر آن است که با استفاده از داده‌های محل قرارگیری خودپردازهای بانک ملت، محل قرارگیری پایانه‌های غیر نقدی و خودپردازهای غیر ملتی را به صورت طول و عرض جغرافیایی پیدا کرده، سپس با خوشه بندی خودپردازها روی نقشه، مناطقی که نیاز به افزایش، کاهش یا جابجایی خودپرداز دارند مشخص و نهایتا با شناسایی خودپردازهای رقیب، مناطق ناکارآمد بانک ملت، نواحی مستعد نصب خودپرداز |

و البته مختصات اشتباه خودپردازهای بانک ملت تعیین گردند. به همین منظور توجه به ارزندگی مشتریان در آن منطقه، میزان امنیت آن منطقه، هزینه نگهداشت و پول‌گذاری، میزان نزدیکی به شعبه، هزینه‌ها و درآمدهای مبتنی بر کارمزد در آن منطقه، تعداد و مبالغ تراکنش های روز خودپرداز، میزان نیاز به خودپرداز غیر نقدی به جای خودپرداز نقدی و... حائز اهمیت می‌باشد. لازم به ذکر است که به منظور حل این امر تنها داده‌های محل قرارگیری اکثر خودپردازهای ملتی و زمان و مبلغ تراکنش‌های کارت‌های ملتی روی خودپرداز سایر بانک‌ها در دسترس می‌باشند.

روش اصلی حل این مسئله، بررسی شباهت سبد خرید[1]دو خودپرداز می‌باشد که به بررسی شباهت در افراد و فاصله زمانی این تکرار می‌پردازد. از آنجا که در صورت نزدیک بودن دو خودپرداز احتمالا در طول یک سال لااقل دو نفر در فاصله زمانی کمتر از ۵ دقیقه از هر دوی این خودپردازها استفاده کرده‌اند را می‌توان به عنوان ملاک مناسبی در نظر گرفت. پس برای اینکه دو خودپرداز نزدیک یکدیگر باشند لازم است که تعداد شماره کارت‌های منحصربه‌فردی که در فاصله زمانی کوتاه از هر دوی این خودپردازها استفاده کرده‌اند مورد بررسی قرار گیرند و خودپردازهایی که تعداد موارد مشترک زیادی دارند در یک خوشه قرار گیرند. از آنجا که حداقل مختصات جغرافیایی یکی از بانک ملت‌های درون خوشه مشخص است، با میانگین گرفتن از مختصات بانک ملت‌های درون خوشه، این مختصات به سایر خودپردازهای درون خوشه، تعمیم می‌یابد. برای خوشه‌هایی که هیچ خودپرداز ملتی درون آن‌ها نیست، برای هر خودپرداز موجود در آن خوشه یک لیست از خودپردازهایی که با خودپرداز مد نظر لااقل ۱۰ تراکنش با فاصله زمانی زیر یک ساعت دارند تهیه می‌گردد. سپس با اختصاص ضریب بیشتر برای خودپردازهای نزدیکتر (تعداد تراکنش با کارت مشترک با فاصله زمانی کمتر) یک میانگین وزن‌دار از مختصات خودپردازهای مجاور گرفته و به خودپرداز مذکور اختصاص داده می‌شود. در ادامه برای خودپردازهای باقی‌مانده همین روال با در نظر گرفتن مختصاتی که تا کنون کشف شده، انجام می‌شود. در نهایت با تطبیق اطلاعات به دست آمده با اطلاعات ثبت شده در نرم‌افزارهای حاوی اطلاعات نقشه، یافته‌ها مورد اعتبارسنجی قرار می‌گیرند.

**شناسایی داده**

برای آغاز پروژه از دو دسته داده استفاده شده است. دسته اول داده‌هایی است که شامل مختصات جغرافیایی خودپردازهای نقدی و غیر نقدی می‌باشند و دسته دوم شامل داده‌هایی است که در آن لیست تراکنش‌هایی وجود دارد که روی پایانه‌های بانک ملت یا توسط کارت‌های بانک ملت انجام شده‌است.

داده‌های مربوط به مختصات جغرافیایی خودپردازها از روی دیتابیس موجود در شرکت با استفاده از

[1] Collaborative Filtering Algorithm

کوئری‌های مربوطه نوشته شده آماده شد. و در نهایت پس از بررسی صحت داده‌ها جدول ایجاد شده با استفاده از کتابخانه pyodbc بعنوان یک dataframe وارد کد پایتون مربوطه شد و پردازش‌های لازم روی آن انجام شد.

داده‌ها شامل ایراداتی از قبیل عدم تطبیق نوع تراکنش با واقعیت، عدم تطابق مختصات مذکور برای خودپردازهای بانک ملت با واقعیت، عدم دسترسی به مختصات برخی از پایانه‌های بانک ملت و ... بود که برای هر یک راهکار مختص به خود، به شرح ذیل ارائه گردید:

• عدم تطبیق نوع تراکنش ذکر شده در داده‌ها با واقعیت: برای این منظور با انجام یک سری تراکنش دانسته و مشخص، به ازای هر کد نوع تراکنش حقیقی مشخص شده و بر اساس آن‌ها اصلاحات صورت پذیرفت.

• عدم تطابق مختصات مذکور برای خودپردازهای بانک ملت با واقعیت: از آنجا که مهمترین اطلاعات برای کشف مختصات سایر خودپردازها، مطابق روش ذکر شده در بالا، دانستن مختصات خودپردازهای بانک ملت است، صحت این داده‌ها از ملزومات انجام پروژه است. لیکن پس از بررسی‌های متعدد دریافت شد که مختصات داده شده انطباق کامل با واقعیت ندارند.

**الگوریتم ها و تکنیک ها**

الگوریتم استفاده شده در این روش یک الگوریتم ابتکاری است که مهمترین فاکتور مورد بررسی برای کشف مختصات هر خودپرداز ناشناخته، فاصله آن تا یک خودپرداز شناخته شده است. برای یافتن این فاصله به بررسی کارت‌های ملتی که در فاصله زمانی کوتاه به هر دوی این خودپردازها مراجعه کرده‌اند پرداخته شده است. از آنجا که در صورت استفاده از یک کارت مشخص با فاصله زمانی کوتاه در دو خودپرداز حتما این دو خودپرداز در فاصله جغرافیایی قرار دارند که می‌توان در فاصله زمانی بسیار کوتاه از یکی به دیگری نقل مکان کرد، پس این دو خودپرداز در نزدیکی یکدیگر قرار دارند و با سایر خودپردازهای مجاور یک کلونی از خودپردازها را ایجاد می‌کنند. در این روش با خوشه بندی خودپردازها با فاصله زمانی دو تراکنش کوتاه در یک خوشه قرار گرفتند و مختصات جغرافیایی خودپرداز بانک ملتی به تمامی خودپردازهای آن خوشه تعمیم داده شد. اما از آنجا که نمی‌توان برای تمام خودپردازهای بانک‌ها در سراسر کشور یک خودپرداز ملتی مشخص در نظر گرفت، با مشکل کمبود خودپردازهای ملتی دارای تراکنش مشترک زیر ۲ دقیقه با خودپرداز غیرملتی مواجه شده که جهت حل این مشکل برای مواردی که در فاصله زیر ۲ دقیقه هیچ تراکنش با کارت یکتای مشترک با خودپرداز بانک ملت نداشتند یک تطابق و میانگین وزن‌دار به ثقل تعداد تراکنش مشترک در مدت زمان یک سال بر روی مختصات جغرافیایی بانک‌های ملت مجاور صورت پذیرفت. بدین ترتیب که از آنجا که وقتی دو خودپرداز به یکدیگر نزدیکتر هستند، احتمالا تراکنش مشترک بیشتری

دارند. پس مختصات جغرافیایی نهایی با توجه به تعداد تراکنش مشترک در زمان زیر یک ساعت ضریب داده شد که باعث گردید مختصات خودپرداز مجهول شبیه‌تر به مختصات خودپردازی شود که بیشترین تعداد تراکنش با یک کارت یکسان با فاصله زمانی کوتاه را داشته‌است. نهایتا مختصات خروجی یک برآیندی منطقی از خودپردازهای مجاور خودپرداز اصلی است. همچنین در بررسی‌های انجام شده صحت و درستی این روش به طور کامل مورد تایید است که در ذیل به بررسی آن پراخته می‌شود.

| **معیار(بنچمارک)** |
| --- |
| یکی از روش‌های بررسی صحت نتایج، بررسی‌های میدانی و بررسی با استفاده از نرم‌افزارهای جِئوگرافیکی مانند Google Map و ... است که با نمونه‌گیری‌های تصادفی از نتایج، درستی آن‌ها مورد بررسی قرار می گیرد.<br>همچنین با محاسبه و تخمین خودپردازهای ملتی از روی سایر خودپردازها با همان روش به صحت و درستی الگوریتم و اجرای آن موقن شد. |

### ۲-۲-۲ شناسایی و کشف رفتارهای نامتعارف تراکنش‌ها

یکی دیگر از پروژه‌ها، شناسایی نوع تراکنش‌ها بر اساس رفتار تراکنشی و شناسایی تراکنش‌های مشکوک به تقلب برای شرکت به‌پرداخت ملت بود. در این پروژه، الگوریتم‌های یادگیری ماشین برای تحلیل داده‌های تراکنش‌ها و شناسایی الگوهای رفتاری مشکوک توسعه داده شد. این پروژه به بهبود امنیت تراکنش‌های بانکی و کاهش ریسک‌های مرتبط با تقلب کمک کرد. . این پروژه در طول مدت زمان کارآموزی اینجانب به پایان نرسید، اما در طول این مدت استخراج دادگان و طراحی مدل های اولیه برای این پروژه صورت پذیرفت. در حال حاضر در شرکت در حال کارکرد بر روی این پروژه هستیم.

### ۲-۲-۳ تحلیل حواله‌های کارت به کارت

در این پروژه، به تحلیل حواله‌های کارت به کارت با صادرکنندگی بانک ملت پرداخته شد. هدف شناسایی مشتریان مستعد پرداخت کارمزد بود که با تحلیل الگوهای تراکنش و رفتار مشتریان، مدل‌هایی برای شناسایی این مشتریان توسعه داده شد. . این پروژه نیز در طول مدت زمان کارآموزی اینجانب به پایان نرسید، اما در طول این مدت استخراج دادگان و طراحی مدل های اولیه برای این پروژه صورت پذیرفت. در حال حاضر در شرکت در حال کارکرد بر روی این پروژه هستیم.

### ۴-۲-۲ پیش‌بینی رفتار مشتریان به منظور ارائه یک سیستم هشدار سریع جهت پیش‌بینی خروج مشتریان از "همراه بانک ملت" (پروژه Churn)

در این پروژه، هدف پیش‌بینی خروج مشتریان از اپلیکیشن "همراه بانک ملت" بود. با استفاده از تکنیک‌های یادگیری ماشین و تحلیل داده‌های کاربری، مدلی برای پیش‌بینی رفتار مشتریان و احتمال ترک اپلیکیشن توسط آن‌ها توسعه داده شد. این مدل به بانک ملت کمک کرد تا اقدامات مناسبی برای حفظ مشتریان و افزایش رضایت آن‌ها انجام دهد.

| **بیان مسأله پروژه** |
| --- |
| در این پروژه هدف اصلی شناسایی و جلوگیری از رویگردانی مشتریانی است که کارمزد بالایی بابت تراکنش‌های خود در بستر "همراه بانک ملت" پرداخت می‌کنند. این طرح به دنبال توسعه مدلی است که با دقت مناسب قادر به پیش‌بینی رفتار مشتریان و شناسایی مشتریان مستعد ریزش باشد تا بتوان اقدامات لازم برای حفظ این دسته از مشتریان را به‌موقع انجام داد.<br>از جمله منافع این پروژه می‌توان به موارد زیر اشاره کرد:<br>- خوشه‌بندی مشتریان بر اساس رفتار تراکنش آنها در بستر همراه بانک<br>- شناسایی مشتریان فعال و وفادار و همچنین مشتریان مستعد ریزش<br>- تفکیک مشتریان بر اساس حجم تراکنش‌ها و کارمزدهای پرداختی آنها<br>با توجه به اهمیت این مساله، ابتدا نیاز به تحلیل رفتار تراکنشی مشتریان است تا دلیل رویگردانی آنها از همراه بانک شناسایی و رفع شود. برای این منظور، استفاده از یک مدل دقیق و کارا برای مدیریت و کاهش ریزش مشتریان ضروری است. در این پروژه از داده‌های تراکنش‌های مشتریان در بستر همراه بانک و سایر کانال‌ها استفاده می‌شود تا برچسب‌گذاری مشتریان به عنوان فعال یا غیر فعال انجام گیرد. مدل‌سازی بر اساس متغیرهایی نظیر تعداد تراکنش‌ها، مبلغ کارمزد و تعداد روزهای فعال بودن مشتری در همراه بانک صورت می‌گیرد. |

| **شناسائی داده** |
| --- |
| برای شروع این پروژه، داده‌های مورد استفاده شامل تراکنش‌های موبایل‌بانک و اطلاعات مربوط به مشتریان آن بوده است. در این بخش، ابتدا مشتریان بر اساس تعداد تراکنش‌های ماهانه‌شان در بستر موبایل‌بانک، با استفاده از سه حد آستانه مختلف (۲، ۵ و ۸ تراکنش) برچسب‌گذاری شدند. |

برای هر یک از این حدود آستانه، مشتریانی که تعداد تراکنش آن‌ها کمتر از مقدار تعیین‌شده بود، به عنوان مشتریان "غیرفعال" و مشتریانی که تعداد تراکنش آن‌ها بیشتر یا مساوی حد آستانه بود، به عنوان مشتریان "فعال" دسته‌بندی شدند. پس از انجام عملیات پیش‌پردازش شامل حذف مقادیر null و duplicate و همچنین متوازن کردن داده‌ها، برای پیش‌بینی این دسته‌ها مدل‌های مختلف یادگیری ماشین شامل Logistic Regression، Decision Tree، Random Forest، XGBoost و SVM توسعه و اجرا شدند.

همچنین در ادامه، متغیر دیگری به نام مبلغ کارمزد تراکنش‌ها نیز برای برچسب‌گذاری و پیش‌بینی مشتریان فعال و غیرفعال مورد استفاده قرار گرفت. مشتریانی که مبلغ کارمزدشان در ماه آخر کمتر از ۱۰،۰۰۰ ریال بود، به عنوان "غیرفعال" و مشتریانی که مبلغ کارمزدشان بیشتر یا مساوی این مقدار بود، به عنوان "فعال" دسته‌بندی شدند.

نتایج حاصل از اجرای این مدل‌ها برای هر یک از حدود آستانه و بر اساس معیارهای مختلفی از جمله Accuracy، Precision، Recall، F-measure و ROC، ارائه شده است که در ادامه این گزارش به تفکیک هر معیار مورد بررسی قرار گرفته‌اند.

**الگوریتم ها و تکنیک ها**

در این پروژه، به منظور پیش‌بینی ریزش مشتریان موبایل بانک و جلوگیری از این رخداد، مراحل مختلفی با استفاده از الگوریتم‌های متنوع یادگیری ماشین و شبکه‌های عصبی انجام شد. در ادامه به تشریح دقیق‌تر هر یک از مراحل و الگوریتم‌های به کار گرفته شده پرداخته می‌شود:

۱. برچسب‌گذاری مشتریان بر اساس حد آستانه :

در این مرحله، مشتریان به دو دسته "فعال" و "غیرفعال" بر اساس تعداد تراکنش‌هایشان طبقه‌بندی شدند. سه حد آستانه ۲، ۵ و ۸ برای تعداد تراکنش‌های همراه‌بانک مشتری برای این کار تعیین شد:

حد آستانه ۲: مشتریانی که تعداد تراکنش‌هایشان کمتر از ۲ بود به عنوان "غیرفعال" و مشتریانی که بیشتر از ۲ تراکنش داشتند به عنوان "فعال" شناخته شدند.

حد آستانه ۵ و ۸: به همین ترتیب، از این آستانه‌ها برای تفکیک مشتریان به دو دسته استفاده شد.

در این مرحله، پیش‌پردازش داده‌ها شامل حذف مقادیر گمشده (null) و داده‌های تکراری (duplicate) انجام شد. سپس مدل‌های یادگیری ماشین شامل رگرسیون لجستیک (Logistic Regression)، درخت تصمیم (Tree Decision)، جنگل تصادفی (Random Forest)، XGBoost و ماشین بردار پشتیبانی (SVM) توسعه و ارزیابی شدند. عملکرد این مدل‌ها بر اساس معیارهای دقت (Accuracy)، دقت مثبت (Precision)،

حساسیت (Recall) و معیار F (F-Measure) سنجیده شد که نشان‌دهنده عملکرد مشابه مدل‌ها در این معیارها بود. که دقت مدل ها برای حد آستانه ۸ بعنوان نمونه در ادامه ذکر خواهد شد.

۲. پیش‌بینی تعداد تراکنش‌ها و کارمزد با استفاده از مدل‌های رگرسیون و شبکه عصبی:

در این مرحله، هدف پیش‌بینی تعداد تراکنش‌های مشتریان برای دوره‌های آتی بود. برای این منظور از مدل‌های رگرسیون و شبکه‌های عصبی استفاده شد:

مدل‌های رگرسیون: شامل رگرسیون خطی (Linear Regression)، درخت تصمیم (Decision Tree)، جنگل تصادفی (Random Forest و XGBoost) برای پیش‌بینی تعداد تراکنش‌ها و کارمزد به کار گرفته شدند. متغیر پاسخ در این مدل‌ها، میانگین تعداد تراکنش‌ها و مبلغ کارمزد در سه ماه منتهی به آذر انتخاب شد.

مدل شبکه عصبی LSTM : این مدل نوعی از شبکه‌های عصبی بازگشتی (RNN) است که به دلیل قابلیت یادآوری بلندمدت خود، برای پیش‌بینی داده‌های سری زمانی مانند تراکنش‌های مشتریان به کار می‌رود. LSTM با استفاده از ساختار خاص خود، قادر به تحلیل الگوهای زمانی و پیش‌بینی تعداد تراکنش‌های دوره بعدی است.

با استفاده از بکارگیری یکی از الگوریتم‌های شبکه عصبی موسوم به LSTM[۲] می‌باشد. شبکه LSTM نوع خاصی از شبکه RNN[۳] است که مشکل حافظه بلندمدت شبکه RNN را حل می‌کند. این شبکه سازوکارهایی داخلی به اسم گیت[۴]دارد که جریان اطلاعات را کنترل می‌کنند؛ این گیت‌ها همین‌طور مشخص می‌کنند چه داده‌هایی در توالی حائز اهمیت هستند و باید هم‌چنان حفظ شوند و چه داده‌هایی باید حذف شوند؛ به این شکل، شبکه اطلاعات مهم را در طول زنجیره توالی عبور می‌دهد تا خروجی مدنظر را ارائه دهد.

در مساله پروژه پیش‌بینی تعداد تراکنش مشتریان موبایل بانک، با توجه به اینکه تعداد تراکنش هر مشتری توالی از اعداد در طول زمان می‌باشد می‌توان با کمک مدل LSTM و بکارگیری پنجره زمانی مورد نظر، تعداد تراکنش دوره بعدی هر مشتری را محاسبه نمود.

در این بخش با اجرای مدل LSTM بر روی دیتاست مورد نظر، تعداد تراکنش مشتریان موبایل بانک برای دوره یا ماه آتی با توجه به پنجره زمانی تعریف شده (یک سال قبل) پیش‌بینی می‌شود.

[2] Long-Short Term Memory
[3] Recurrent Neural Network
[4] Gate

به منظور اجرای مدل LSTM، ابتدا برخی عملیات پیش پردازش مانند Scaling، تبدیل متغیرهای کیفی[5] به متغیرهای مجازی صفر و یک و سپس برخی عملیات Feature Engineering مانند تعریف پنجره زمانی، اضافه کردن برخی متغیرهای جدید با توجه به متغیرهای فعلی و تعریف متغیر پاسخ (لیبل) صورت پذیرفت.

پس از انجام مراحل پیش پردازش، در گام اول به منظور در نظر گرفتن پنجره زمانی یازده ماهه از هر مشتری برای پیش‌بینی تراکنش ماه بعدی، مشتریانی انتخاب شدند که در بیش از دوازده ماه تراکنش داشتند. با این کار تعدادی مشتری فیلتر و انتخاب شدند.

پس از انجام عملیات پیش‌پردازش و آماده‌سازی دیتا، قدم اول در توسعه مدل‌های یادگیری عمیق این است که لایه‌های شبکه عصبی تعریف و ساخته شوند، هر لایه شامل نورون‌هایی است که در آن تعداد لایه‌ها و نورون‌ها توسط تحلیلگر تعریف می‌شود. سپس با استفاده از یکی از شیوه‌ها و توابع مرسوم در مدل‌های یادگیری عمیق، لایه‌ها به صورت یک لیست تعریف و به تابع Sequential داده می‌شوند، به گونه‌ای که هر یک از درایه‌های این لیست یک لایه شبکه عصبی محسوب می‌شوند. برای ساخت لایه‌ها از داخل کتابخانه Keras لایه‌های LSTM و Dense و عملگر Dropout به مدل اضافه می‌شوند. پس از ساختن لایه‌ها گام بعدی در توسعه مدل‌های شبکه عصبی، Compile کردن مدل است که خود شامل سه آرگومان است.

اولین آرگومان از این گام تعریف الگوریتمی جهت حل مساله بهینه‌سازی است. عموما مسائل شبکه عصبی، مسائل Non-convex (غیر محدب) محسوب می‌شوند که لزوما دارای یک نقطه بهینه نیستند بلکه دارای چندین نقطه بهینه محلی بوده و ممکن است یک نقطه بهینه سرتاسری یا مطلق داشته باشند. پایه و اساس تمام این الگوریتم‌ها، الگوریتم گردایان کاهشی[6] می‌باشد که الگوریتمی تکرار شونده بوده و بر مبنای مشتق کار می‌کند. از این الگوریتم برای بدست آوردن وزن‌ها و بایاس‌ها (پارامترهای موجود در شبکه عصبی) استفاده می‌شود. به واسطه این گام پروسه یادگیری[7] در مدل شبکه عصبی انجام می‌شود.

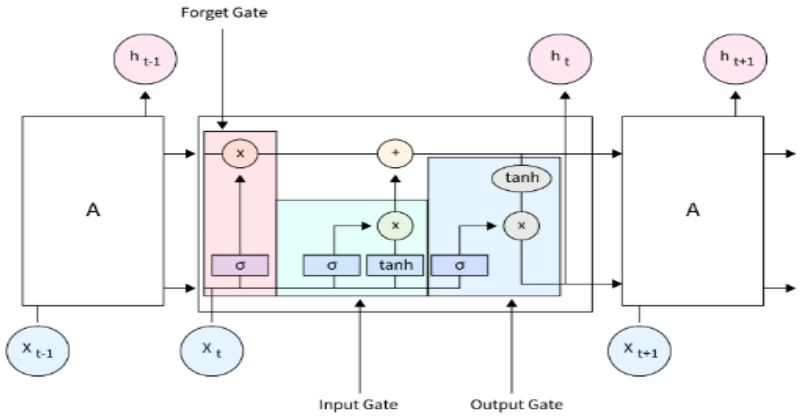
آرگومان دوم از گام Compile کردن مدل، افزودن تابع ضرر[8] مربوطه می‌باشد که وظیفه آن محاسبه خطای پیش‌بینی از طریق محاسبه گرادیان می‌باشد.

آرگومان سوم مربوط به آرگومان metric یا شاخص است که با توجه به جنس متغیر پاسخ اعم از کیفی و کمی، می‌توان از شاخص‌های مختلفی استفاده کرد اما لزومی برای تعریف این آرگومان وجود ندارد و اضافه کردن آن تاثیر خاصی در مساله بهینه‌سازی ندارد و فقط با تعریف آن مشخص می‌کنیم که هنگام خروجی گرفتن از مدل این شاخص نمایش داده شود.

[5] Categorical
[6] Gradient descent
[7] Learning
[8] Loss Function

پس از تعریف لایه‌ها و compile کردن، نوبت به برازش مدل می‌رسد که این بخش نیز دارای چند آرگومان مختلف و مهم می‌باشد. به عبارتی با اعمال تابع برازش بر روی دیتای آموزشی، تمام پارامترها از جمله وزن‌ها بدست می‌آیند.

برخی از آرگومان‌های تابع برازش در مدل‌های شبکه عصبی، شامل epoche و batch_size می‌باشد که در آن هر epoch شامل کل مشاهداتی است که به الگوریتم جهت آموزش داده می‌شود. آرگومان batch_size نیز نماینده دسته‌هایی از مشاهدات است که به ازای آنها هر بار پارامترها آپدیت می‌شوند. برای مثال در هر epoch اگر batch_size مقدار ۳۰ در نظر گرفته شود و کل مشاهدات ۶۰۰۰ رکورد باشد، در هر batch هر بار بر روی تعداد ۲۰۰ مشاهده مشتق حساب شده و پارامترها آپدیت می‌شوند به عبارتی در هر epoch، مشاهدات به تعداد ۳۰ بار در دسته‌های ۲۰۰ تایی آپدیت می‌شوند.




تصویر ۱-۱ نحوه عملکرد کلی و ساختار شبکه عصبی بازگشتی LSTM

۳. استفاده از تکنیک‌های بدون نظارت برای خوشه‌بندی مشتریان:

در این فاز، از روش‌های یادگیری بدون نظارت برای کشف الگوهای رفتاری مشتریان و شناسایی دسته‌های مستعد ریزش استفاده شد:

با توجه به اینکه در مساله ریزش مشتریان موبایل بانک، داده‌های مورد استفاده برای تحلیل بدون برچسب هستند، در این فاز به منظور کشف الگوها و اطلاعاتی که در داده‌ها وجود دارد از تکنیک یادگیری بدون نظارت[9] به عنوان یکی از تکنیک‌های یادگیری ماشین استفاده شده است که در آنها خود مدل به تنهایی جهت کشف الگوها و اطلاعاتی که در داده وجود دارد کافی می‌باشد. در مرحله اول این فاز برای تعیین تعداد خوشه بهینه از شاخص Elbow استفاده شده است. پس از تعیین تعداد خوشه بهینه از روش K-Means جهت خوشه‌بندی کلان داده‌ها از نرم‌افزار SPARK استفاده شده است. در مرحله دوم برچسب‌گذاری

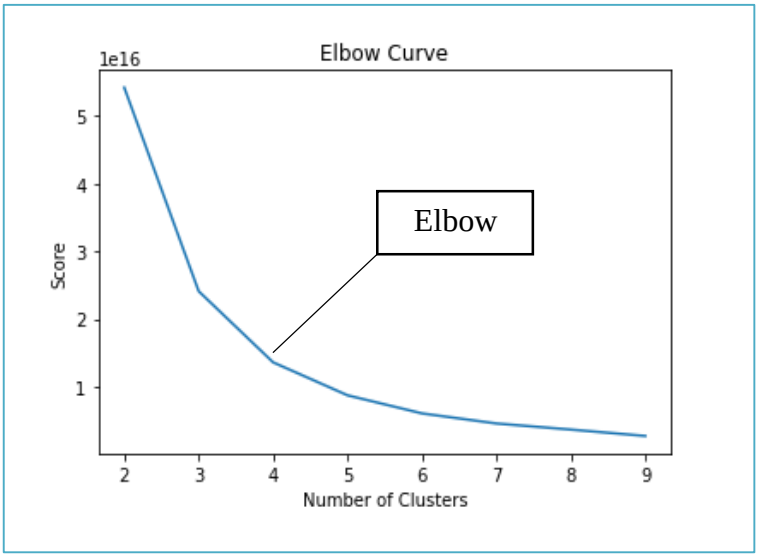
[9] Unsupervised Learning

مشتریان برای تمامی خوشه‌های بدست آمده در مرحله قبلی انجام شده و در نهایت نتایج حاصله مورد ارزیابی قرار گرفته و پیشنهادهایی جهت ارتقاء عملکرد کسب وکار موبایل بانک ذکر شده است.

از بین تمام متغیرهای موجود در دیتاست جمع‌آوری شده، متغیرهایی که می‌توانستند رفتار تراکنشی مشتریان در موبایل را به خوبی توصیف کنند کافی بودند. بنابراین ابتدا با توجه به متغیر "تعداد تراکنش‌های صورت گرفته در بستر موبایل بانک" مشتریان را خوشه‌بندی کرده و سپس با کمک گرفتن از همبستگی این متغیر با متغیر "تعداد تراکنش‌های صورت گرفته در کانال‌های پتانسیل‌دار" و متغیر Threshold، مشتریان برچسب‌گذاری شدند.

**گام اول-خوشه‌بندی مشتریان:**

به منظور رسیدن به تفکیک منطقی از مشتریان با در نظر گرفتن رفتار تراکنشی آنها در بستر موبایل بانک و رصد مشتریانی که دارای رفتار رویگردانی از موبایل بانک هستند، اقدام به خوشه‌بندی مشتریان با استفاده از روش خوشه‌بندی K-Means و معیار مشابهت کوسینوسی[1] بر اساس متغیر تراکنش‌های صورت گرفته در موبایل بانک شد. در این روش هر یک از مشتریان به گونه‌ای خوشه‌بندی شدند که بیشترین مشابهت را به مشاهدات درون خوشه خود و کمترین مشابهت را به مشاهدات خوشه‌های دیگر داشته باشند. حساس بودن




تصویر ۰۱-۲ نمایی از منحنی آرنجی در الگوریتم خوشه بندی

این الگوریتم به مراکز خوشه اولیه سبب می‌شود که تنها بتوان یک پاسخ بهینه محلی تولید کرد.

یکی از الزامات شروع خوشه‌بندی، تعیین تعداد خوشه‌ها از طریق یکی از روش‌های تعیین تعداد خوشه بهینه است. در این مساله از روش Elbow به عنوان یکی از این روش‌ها استفاده شده است. این روش درصد واریانس

[1] Cosine Similarity

را به عنوان تابعی از تعداد خوشه‌ها در نظر می‌گیرد و تعداد خوشه‌ها را به گونه‌ای انتخاب می‌کند که با اضافه‌کردن خوشه‌ای دیگر، مدل‌سازی داده بهتری بدست نیاید. در زیر تعداد خوشه‌های بهینه جهت تفکیک مشتریان بر اساس رفتار تراکنشی آنها در بستر موبایل بانک نشان داده شده است.

با توجه به نمودار فوق مشخص است که بعد از نقطه ۴ که نشاندهنده تعداد خوشه‌ها می‌باشد، کاهش قابل توجهی در میزان واریانس موجود در دیتا مشاهده نمی‌شود، بنابراین تعداد چهار خوشه به عنوان تعداد خوشه بهینه انتخاب می‌گردد.

هدف از این خوشه‌بندی رسیدن به تفکیک و تقسیم‌بندی مشتریانی است که دارای رفتار تراکنشی مشابهی در بستر موبایل بانک هستند. نکته مهم این است که به دلیل پویا و احتمالی بودن تعداد تراکنش مشتریان در بستر موبایل بانک و اینکه پنجره زمانی محدودی از رفتار مشتریان در اختیار ماست، آینده آنها دور از دسترس دید بوده و نمی‌توان با قطعیت در مورد روند الگوی رفتاری آنها برای دوره‌های آتی قضاوت کرد. زیرا در برخی موارد روند سری زمانی مربوط به تراکنش‌های صورت گرفته در موبایل بانک از الگوی خاصی پیروی نکرده و این سری زمانی تابعی از زمان نمی‌باشد.

جهت اثبات این موضوع بر روی نمونه‌ای از سری زمانی تعداد تراکنش برخی از مشتریان در بستر موبایل بانک، آزمون فرضی صورت گرفت که در آن نشان داده شده است که لزوما سری زمانی مربوط به تراکنش‌های موبایل بانک همه مشتریان دارای ساختاری تابع از زمان نیستند. به این منظور با در نظر گرفتن فرض صفری مبنی بر اینکه سری زمانی تعداد تراکنش مشتریان موبایل بانک دارای ساختاری وابسته به زمان می‌باشد؛ آزمون فرضی با استفاده از آماره آزمون دیکی فولر افزوده[1] انجام شد. پس از انجام این آزمون با توجه به مقدار p-value، فرض صفر رد و مشخص گردید سری زمانی تعداد تراکنش برخی از مشتریان دارای ساختار وابسته به زمان نمی‌باشد. به عبارت دیگر این سری‌ها توسط الگو یا روند خاصی که وابسته به زمان باشند، قابل تعریف نیستند.

از این رو بهتر است برای تفکیک مشتریان مستعد رویگردانی با فرض عدم وجود روند یا الگوی خاص در سری زمانی ثبت شده برای تعداد تراکنش آنها در بستر موبایل بانک، به دنبال یافتن مشتریانی باشیم که در یک مقطع زمانی ایستا با کاهش تراکنش‌های خود در بستر موبایل بانک، آنها را به کانال‌هایی انتقال می‌دهند که قابلیت انجام در موبایل بانک را دارند. به این معنی که در بین خوشه‌های مشتریان، مشتریانی وجود دارند که به هر دلیلی در حال جابجایی تراکنش‌های موبایل بانک خود به بسترهای غیر از موبایل بانک هستند و این در حالی است که تراکنش‌های مربوطه قابلیت و پتانسیل انجام در موبایل بانک را نیز دارند. از این رو با رصد این مشتریان می‌توان به دسته‌ای از مشتریانی رسید که در حال کاهش علاقه‌مندی به انجام تراکنش

[1] Augmented Dickey Fuller Test [1]

در بستر موبایل بانک هستند. سرانجام با جویا شدن علت انتقال تراکنش‌های موبایل بانک به دیگر بسترها می‌توان به نقاط ضعف و نقص‌های موجود در عناصر مربوط به کسب و کار موبایل بانک پی برده و در صدد رفع آنها و در نتیجه حفظ و رضایت‌مندی مشتری و به تبع آن کسب سود برآمد.

نتایج حاصل از خوشه‌بندی مشتریان با توجه به متغیر تعداد تراکنش در بستر موبایل بانک چهار خوشه می‌باشد. هر کدام از این خوشه‌ها دارای ویژگی آماری و الگوی رفتاری خاصی هستند.

پس از خوشه‌بندی دیتاست مورد نظر، آنچه از بررسی خروجی خوشه‌بندی نتیجه گرفته می‌شود این است که مشتریان خوشه شماره صفر منعکس‌کننده افرادی با کمترین میانگین تعداد تراکنش و بالاترین بی‌ثباتی در تراکنش هستند. این بی ثباتی رفتار به معنی نداشتن الگوی رفتاری مشخص در طول زمان می‌باشد. مشتریان خوشه شماره یک بیانگر افرادی هستند که میانگین تعداد تراکنش و ثبات تراکنشی بالاتری نسبت به خوشه صفر دارند. میانگین تعداد تراکنش در خوشه شماره دو حدود ۸ برابر میانگین تعداد تراکنش مشتریان خوشه شماره صفر و ۲/۵ برابر میانگین تعداد تراکنش خوشه شماره یک بوده و همچنین نسبت به این دو خوشه دارای ثبات تراکنشی بالاتری می‌باشد. خوشه شماره سه خوشه‌ای است که میانگین تعداد تراکنش بالاتری نسبت به سه خوشه دیگر داشته و از ثبات نسبی بالاتری نسبت به این خوشه‌ها برخوردار است.

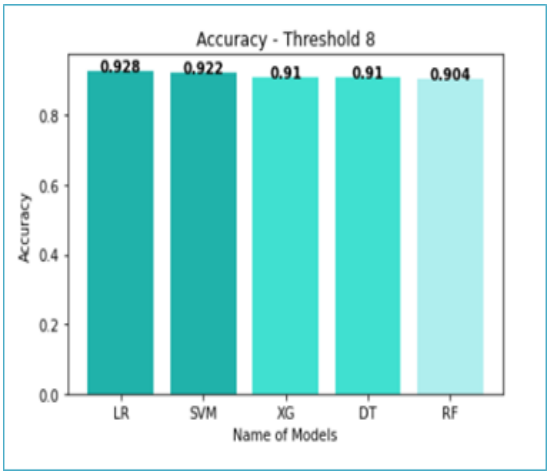
۴. ارزیابی و محاسبه شاخص‌های عملکردی مدل‌ها:

پس از اجرای مدل‌های مختلف، معیارهای RMSE (Root Mean Square Error) برای سنجش دقت مدل‌های رگرسیون و ارزیابی خطاهای پیش‌بینی محاسبه شدند. همچنین، از معیارهای False Negative و False Positive برای ارزیابی تعداد مشتریانی که مدل به اشتباه ریزشی یا غیرریزشی تشخیص داده است، استفاده شد. این ارزیابی‌ها به بهبود مدل‌ها و افزایش دقت پیش‌بینی‌ها کمک کرد.

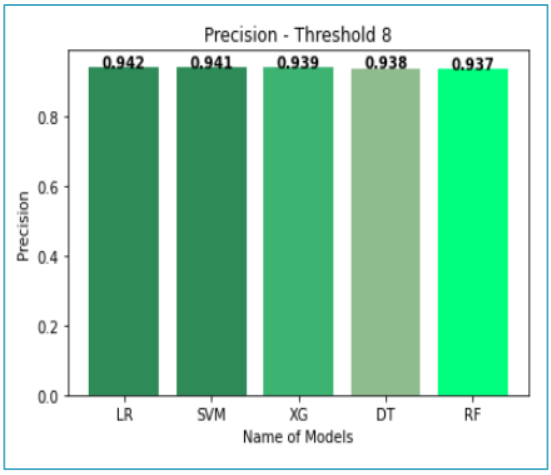
۵. شناسایی مشتریان با تراکنش‌های کارمزددار:

در نهایت، با تحلیل تراکنش‌های موبایل بانک و بررسی میزان کارمزد پرداختی مشتریان، تلاش شد تا مشتریانی که سودآوری بالایی برای بانک دارند و مستعد ریزش هستند، شناسایی شوند. الگوی رفتاری تراکنش‌ها و کارمزدهای پرداختی مشتریان نشان داد که این دو به شکل مشابهی توزیع شده‌اند و این اطلاعات به حفظ مشتریان ارزشمند کمک کرد.

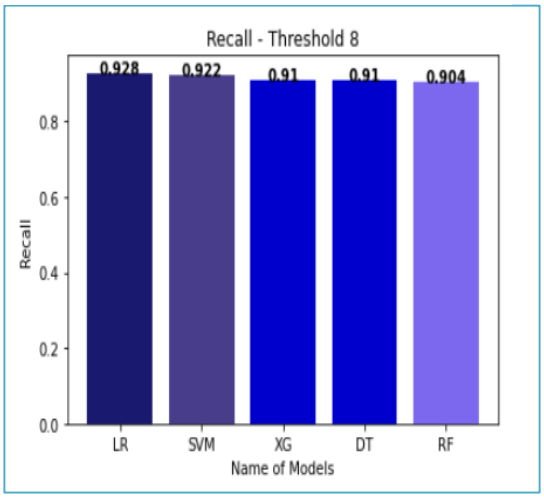
**معیار(بنچمارک) و نتایج حاصله**



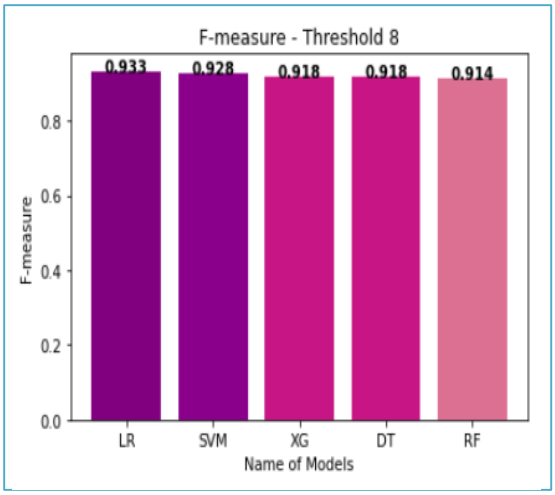

تصویر ۰-۴ نتایج معیار Accuracy به تفکیک هر مدل




تصویر ۰-۳ نتایج معیار Precision به تفکیک هر مدل:



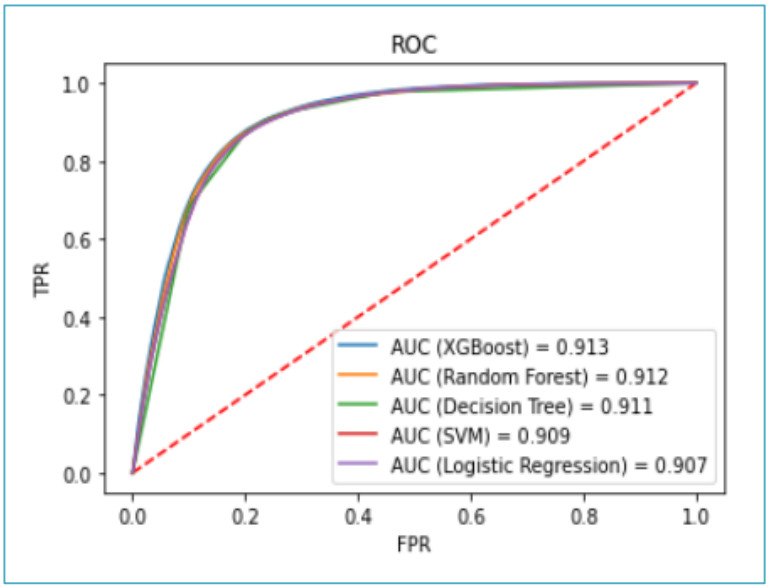

تصویر ۱-۶ نتایج معیار Recall به تفکیک هر مدل




تصویر ۱-۰۵ نتایج معیار F-Measure به تفکیک هر مدل

مدل های ساخته شده برای سایر مقادیر Thereshold و نیز مقدار آستانه مربوط به میزان کارمزد نیز ساخته و تست شد که نتایج آن مشابه با همین نتایج بوده و از ذکر تک تک آن‌ها صرف نظر می‌کنیم.

نتایج حاصل از معیار ROC به تفکیک هر مدل:




تصویر ۰-۷ نمودار ROC برای هر یک از مدل ها

نتایج مقایسه مدل‌ها با توجه به مقدار آستانه ۸ مشابه سایر حدود آستانه بررسی شده حاکی از آن است که مقادیر معیارهای ارزیابی در مدل‌های مختلف از اختلاف بسیار کمی نسبت به هم برخوردار، به گونه‌ای که:

- بر اساس معیار Accuracy، مدل Logistic Regression با مقدار ۰/۹۲۸ بهترین مدل شناخته شد.
- بر اساس معیار Precision، مدل Logistic Regression با مقدار ۰/۹۴۲ بهترین مدل شناخته شد.
- بر اساس معیار Recall، مدل Logistic Regression با مقدار ۰/۹۲۸ بهترین مدل شناخته شد.
- بر اساس معیار F_measure، مدل Logistic Regression با مقدار ۰/۹۳۳ بهترین مدل شناخته شد.
- بر اساس نمودار ROC، مدل XGBoost با مقدار ۰/۹۱۳ بهترین مدل شناخته شد.

در گام بعدی همان طور که ذکر شد با توجه به نیازمندی موجود از مدل های رگرسیونی استفاده شد تا دقیقا پیش بینی کنیم تعداد دفعات استفاده هر مشتری از همراه بانک ملت به چه تعداد خواهد بود. نتایج مدل های پیاده سازی شده در این بخش مطابق زیر بوده است:

مقایسه مدل‌ها در این فاز به دلیل استفاده از مدل‌های رگرسیون توسط متریک RMSE[1][2] سنجیده می‌شود که بیانگر تفاوت میان مقدار پیش‌بینی شده توسط مدل و مقدار واقعی می‌باشد. هر چقدر این مقدار کمتر باشد نشان‌دهنده برازش بهتر مدل است.

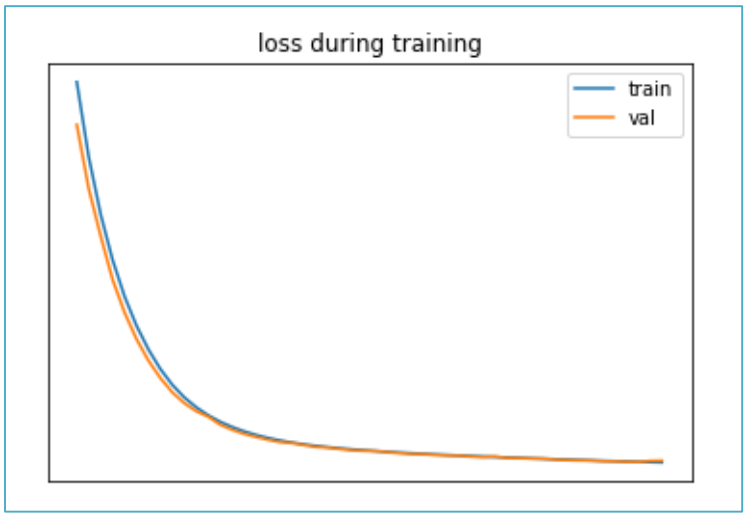
1 Root Mean Square Error 2

جدول ۰-۱ میزان خطا به ازای مدل های یادگیری ماشین برای تسک رگرسیون

| R2Score | MAE | RMSE | نام مدل |
|---|---|---|---|
| 0.44 | 10.81 | 11.29 | Linear Regression |
| 0.61 | 9.95 | 10.93 | XGBoost |
| 0.52 | 10.73 | 11.15 | Random Forest |
| 0.48 | 10.79 | 11.26 | Decision Tree |

با اینکه مطابق جدول فوق مدل XGBoost بیشترین دقت را بین ۴ مدل نهایی دارا بوده است اما این دقت ها کماکان برای مسئله ما کفایت نمی کردند به همین دلیل از یک مدل شبکه عصبی بازگشتی LSTM استفاده شد که نحوه پیاده سازی آن با استفاده از Pyspark پیش‌تر ذکر شد و نتایج آن از قرار زیر است:

با اجرای مدل LSTM میزان خطای مدل آموزش[۳]و مدل اعتبارسنجی[۴]به ترتیب ۰,۲۲ و ۰,۲۳۵ می‌باشد. نمودار حاصل از محاسبه خطا در مدل آموزش و اعتبارسنجی در تصویر زیر نمایش داده شده است. نتایج نشان داده شده در نمودار حاکی از آموزش و برازش مناسب مدل می‌باشد.




تصویر ۰-۸ نمودار میزان خطا به ازای هر ایپاک در مدل LSTM

در نهایت نیز مقادیر RMSE و R2Score و MAE برای این مدل به ترتیب ۴,۱۷ و ۰,۷۹ و ۳,۶۸ می باشد.

[1] Training Loss
[1] Validation Loss
[3]
[4]

### ۲-۲-۵ پیش‌بینی منابع و نقدینگی مشتریان بانک ملت

در این پروژه، هدف پیش‌بینی میزان منابع و نقدینگی مشتریان بانک ملت بود. با تحلیل داده‌های مالی و تراکنش‌های مشتریان، مدلی برای پیش‌بینی میزان نقدینگی و منابع آتی مشتریان توسعه داده شد. این مدل به بانک کمک کرد تا در مدیریت نقدینگی و تخصیص منابع بهینه عمل کند. این پروژه نیز در طول مدت زمان کارآموزی اینجانب به پایان نرسید، اما در طول این مدت استخراج دادگان و طراحی مدل های اولیه برای این پروژه صورت پذیرفت. در حال حاضر در شرکت در حال کارکرد بر روی این پروژه هستیم.

### ۲-۲-۶ ارتقا مدل امتیازدهی پذیرندگان

در این پروژه، هدف بهبود و ارتقا مدل امتیازدهی پذیرندگان بود. با تحلیل داده‌های تراکنش و ویژگی‌های پذیرندگان، مدل‌های جدیدی برای ارزیابی عملکرد و اعتبار پذیرندگان توسعه داده شد تا بتوان به بهبود فرآیندهای اعتباری و کاهش ریسک‌های مالی کمک کرد. این پروژه نیز در طول مدت زمان کارآموزی اینجانب به پایان نرسید، اما در طول این مدت استخراج دادگان و طراحی مدل های اولیه برای این پروژه صورت پذیرفت. در حال حاضر در شرکت در حال کارکرد بر روی این پروژه هستیم.

## فصل سوم: ارزیابی و تحلیل دانشجو از محل کارآموزی

### ۳-۱ ارزیابی کلی

دوره کارآموزی در شرکت مهندسی صنایع یاس ارغوانی یک تجربه بسیار ارزشمند و آموزنده برای من بود. این دوره به من فرصت داد تا دانش و مهارت‌های علمی خود را در زمینه‌های مختلف علم داده و مدیریت خدمات به صورت عملی به کار بگیرم و با چالش‌های واقعی در صنعت مواجه شوم. همکاری با تیمی حرفه‌ای و مشارکت در پروژه‌های بزرگ بانکی، به من دیدگاه عمیق‌تری نسبت به مسائل صنعتی و راه‌حل‌های آن‌ها داد.

### ۳-۲ تحلیل نقاط قوت و ضعف محل کارآموزی

#### ۳-۲-۱ نقاط قوت:

- **تخصص تیم:** تیم شرکت یاس ارغوانی شامل افرادی با تخصص‌های مختلف و تجربه‌های عمیق در زمینه‌های مختلف علم داده و تکنولوژی‌های مرتبط بود. این تعاملات به من کمک کرد تا با دانش و تجربه‌ی حرفه‌ای آن‌ها آشنا شوم و مهارت‌های خود را در زمینه علم داده تقویت کنم. یادگیری از تجربه‌های عملی تیم، به من این امکان را داد که با روش‌های کاربردی در حل مسائل دنیای واقعی آشنا شوم و به درک عمیق‌تری از چالش‌های موجود در پروژه‌ها برسم. همچنین، این فرصت را داشتم که از نظرات و بازخوردهای اعضای تیم استفاده کنم و به بهبود عملکرد خود بپردازم.

- **پروژه‌های واقعی و صنعتی:** پروژه‌هایی که در این شرکت انجام دادم، مستقیماً مرتبط با نیازهای صنعتی بودند و مرا با چالش‌های واقعی کسب‌وکار آشنا کردند. این پروژه‌ها به من این امکان را دادند که مفاهیم علمی و فنی علم داده را به کار ببرم و به درک عمیق‌تری از نیازهای بازار و مشتریان دست یابم. همچنین، کار بر روی پروژه‌های واقعی به من کمک کرد تا مهارت‌های تحلیلی و حل مسئله خود را تقویت کنم و یاد بگیرم که چگونه می‌توانم به طور مؤثر به نیازهای مشتریان پاسخ دهم.

- **زیرساخت بیگ دیتا و پردازش موازی:** در شرکت، با یک کلاستر بیگ دیتا شامل تعداد زیادی سرور کار می‌کردیم که داده‌ها و پردازش‌ها با قدرت زیادی در این زیرساخت انجام می‌شد. این زیرساخت به ما این امکان را می‌داد که داده‌های حجیم را به سرعت پردازش کنیم و تحلیل‌های پیچیده‌ای انجام دهیم. بخش زیادی از کار ما با استفاده از PySpark بود که یک فریمورک پردازش موازی و توزیع‌شده برای داده‌های حجیم است. PySpark به ما این امکان را می‌دهد که با استفاده از زبان برنامه‌نویسی پایتون، پردازش‌های توزیع‌شده را به سادگی پیاده‌سازی کنیم. همچنین برای ارتباط و کار با دیتابیس‌ها، از Apache Impala استفاده می‌کردیم که امکان انجام پرس‌وجوهای سریع روی داده‌های بزرگ را فراهم می‌کند. این تجربه، تمرین بسیار خوبی در زمینه مدیریت پایگاه داده و کار با داده‌های حجیم (Big Data) بود و به من کمک کرد تا با تکنیک‌های پیشرفته‌تری در این زمینه آشنا شوم.

- **آشنایی با مفاهیم و ابزارهای بیگ دیتا:** کار با Hive، Spark، Yarn، و Hue به من این امکان را داد که به طور عملی با مفاهیم پردازش توزیع‌شده آشنا شوم.

  Hive یک سیستم انبار داده است که امکان اجرای پرس‌وجوهای SQL بر روی داده‌های توزیع‌شده در Hadoop را فراهم می‌کند و به من کمک کرد تا با زبان SQL و نحوه کار با داده‌های توزیع‌شده آشنا شوم.
  Spark نیز یک موتور پردازش داده‌های سریع و توزیع‌شده است که با قابلیت پردازش موازی، امکان انجام تحلیل‌های پیچیده روی داده‌های حجیم را فراهم می‌کند. این ابزار به من کمک کرد تا با الگوریتم‌های مختلف پردازش داده آشنا شوم و نحوه بهینه‌سازی عملکرد آن‌ها را یاد بگیرم.
  Yarn (Yet Another Resource Negotiator) وظیفه مدیریت منابع کلاستر و زمان‌بندی کارها را دارد و به مدیریت بهتر فرآیندهای توزیع‌شده کمک می‌کند. این تجربه به من کمک کرد تا بفهمم چگونه می‌توان منابع را بهینه مدیریت کرد و کارایی سیستم را افزایش داد.

Hue یک رابط کاربری وب است که به کار با ابزارهای بیگ دیتا مثل Hive و Impala کمک می‌کند و امکان اجرای پرس‌وجوها و مدیریت داده‌ها را به سادگی فراهم می‌سازد. این ابزار به من کمک کرد تا به صورت بصری با داده‌ها کار کنم و تحلیل‌های خود را به راحتی انجام دهم.

- **ایجاد ارتباطات حرفه‌ای:** یکی از نقاط قوت کلیدی این تجربه، ارتباط با افرادی بود که از جمله‌ی بهترین‌ها در حوزه‌های بازاریابی، مدیریت خدمات، بیگ دیتا، زیرساخت و هوش مصنوعی بودند. این ارتباطات به من فرصت تبادل اطلاعات و یادگیری از افراد متخصص را داد که برای آینده حرفه‌ای من بسیار ارزشمند بود. این شبکه‌سازی به من کمک کرد تا درک بهتری از روندهای صنعت و نیازهای بازار پیدا کنم و همچنین فرصت‌های شغلی آینده را شناسایی کنم.

- **یادگیری مفاهیم CI/CD و Kubernetes** : فرآیندهای مرتبط با CI/CD (یکپارچه‌سازی و تحویل مداوم) در این شرکت هنوز به طور کامل پیاده‌سازی نشده بودند، اما همین موقعیت باعث شد که با این مفاهیم آشنا شوم و یاد بگیرم که چگونه می‌توان از Kubernetes و MLOps برای بهبود فرآیندهای توسعه و استقرار مدل‌های یادگیری ماشین استفاده کرد.
CI/CD فرآیندهایی هستند که به توسعه‌دهندگان کمک می‌کنند تا به‌طور پیوسته کدها را به روز کرده و به صورت خودکار در محیط‌های تولید مستقر کنند. CI به یکپارچه‌سازی کدها و تست‌های مداوم اشاره دارد و CD به تحویل یا استقرار مداوم کدها. این مفاهیم به من کمک کردند تا درک بهتری از نحوه خودکارسازی فرآیندهای توسعه و بهبود کیفیت کدها پیدا کنم.
Kubernetes یک سیستم مدیریت کانتینرها است که استقرار، مقیاس‌پذیری و مدیریت برنامه‌ها در کانتینرها را به صورت خودکار انجام می‌دهد. این ابزار به من کمک کرد تا با مفاهیم کانتینرization و نحوه مدیریت منابع در محیط‌های توزیع‌شده آشنا شوم.
MLOps به استفاده از مفاهیم DevOps در توسعه و استقرار مدل‌های یادگیری ماشین اشاره دارد و به توسعه‌دهندگان کمک می‌کند تا فرآیند استقرار مدل‌های یادگیری ماشین را خودکار و بهینه کنند. آشنایی با این مفاهیم به من کمک کرد تا درک بهتری از چالش‌های موجود در استقرار مدل‌های یادگیری ماشین پیدا کنم و راه‌حل‌های مؤثری برای آن‌ها ارائه دهم.

**۲-۲-۳ نقاط ضعف:**

- **فشار کاری:** در برخی مواقع، حجم کاری بسیار بالا بود که می‌توانست باعث خستگی و کاهش کیفیت کار شود. این تجربه به من کمک کرد تا یاد بگیرم چگونه تحت فشار و در شرایط کاری پرتنش مدیریت زمان و تمرکز خود را حفظ کنم. همچنین، این فشار به من آموخت که چگونه می‌توانم اولویت‌بندی

کنم و وظایف را به طور مؤثر مدیریت کنم تا از بروز استرس جلوگیری کنم.

- **کمبود منابع آموزشی:** در برخی موارد، منابع آموزشی کافی برای یادگیری تکنیک‌های پیشرفته در دسترس نبود. این کمبود می‌توانست به یادگیری عمیق‌تر و تسلط بر موضوعات خاص کمک کند. با این حال، من با استفاده از منابع آنلاین و تجربیات اعضای تیم، این کمبود را جبران کردم و به یادگیری ادامه دادم. این تجربه به من یاد داد که چگونه می‌توانم از منابع موجود به بهترین نحو استفاده کنم و به دنبال یادگیری خود باشم.

- **عدم تکمیل فرآیند CI/CD** : فرآیند CI/CD با استفاده از Kubernetes و MLOps در این شرکت هنوز کامل نشده بود، که به‌عنوان فرصتی برای من بود تا با این مفاهیم بیشتر آشنا شوم و نحوه استفاده از آن‌ها برای بهبود توسعه و استقرار مدل‌های یادگیری ماشین را یاد بگیرم. این عدم تکمیل به من یادآوری کرد که چگونه می‌توان از فرصت‌ها برای یادگیری و رشد استفاده کرد و به من انگیزه داد تا در این زمینه‌ها بیشتر تحقیق کنم.

این تجربیات به من کمک کرد تا مهارت‌های فنی خود را تقویت کنم.

# مراجع

[1] Profiling local optima in K-means clustering: developing a diagnostic technique

[2] A Shape-based Similarity Measure for Time Series Data with Ensemble Learning

[3] How to check if Time Series Data is Stationary with python

[4] https://en.wikipedia.org/wiki/Covariance

[5] http://ceadserv1.nku.edu/longa//classes/mat385/highlights/highlights6.3.pdf